THE GENETIC CODE: II

A sequel to F. H. C. Crick's article of last October, which discussed how the hereditary material embodies the code for the manufacture of proteins. The nature of the code has now been further elucidated

by Marshall W. Nirenberg

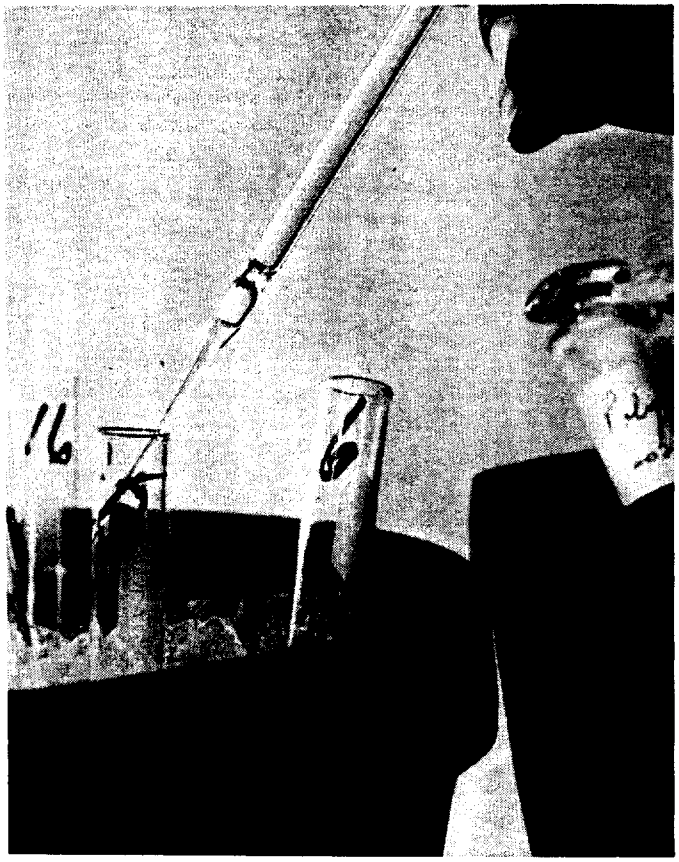
Just 10 years ago James D. Watson and Francis H. C. Crick proposed the now familiar model for the structure of DNA (deoxyribonucleic acid), for which they, together with Maurice H. F. Wilkins, received a Nobel prize last year. DNA is the giant helical molecule that embodies the genetic code of all living organisms. In the October 1962 issue of *Scientific American* Crick described the general nature of this code.

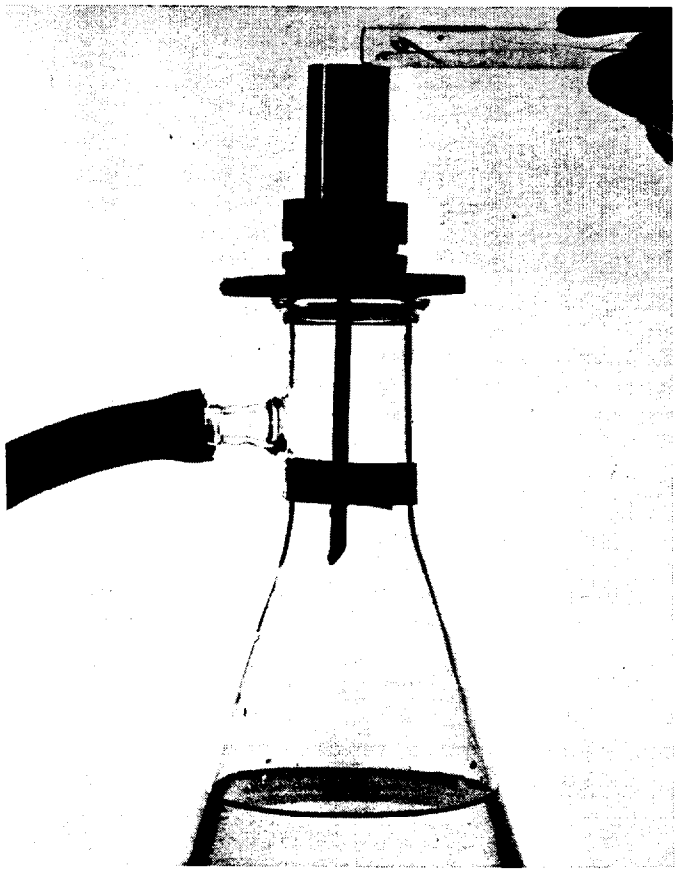
By ingenious experiments with bacterial viruses he and his colleagues established that the "letters" in the code are read off in simple sequence and that "words" in the code most probably consist of groups of three letters. The code letters in the DNA molecule are the four bases, or chemical subunits, adenine, guanine, cytosine and thymine, respectively denoted A, G, C and T.

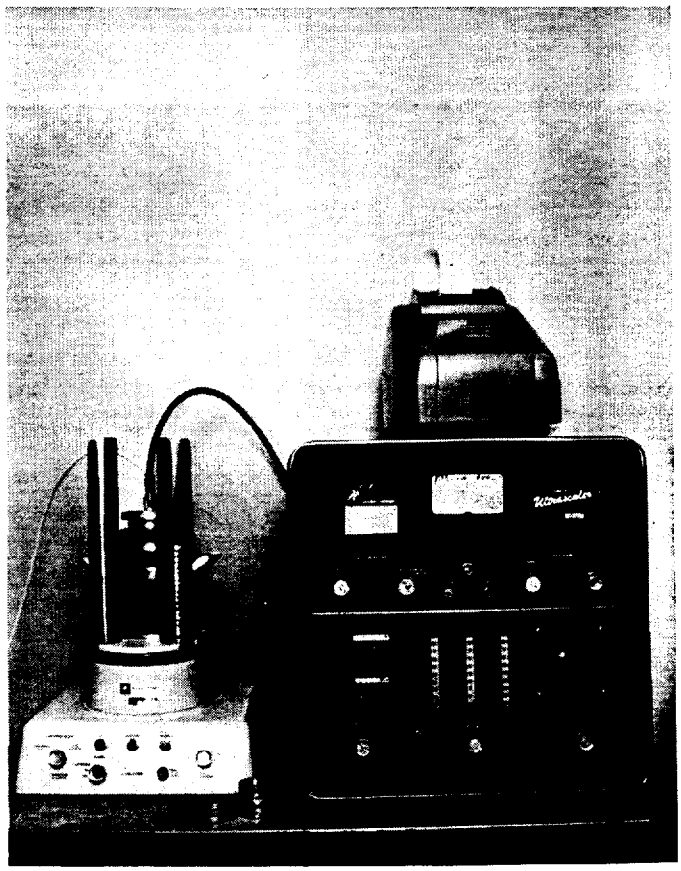
This article describes how various combinations of these bases, or code letters, provide the specific biochemical information used by the cell in the construction of proteins: giant molecules assembled from 20 common kinds of amino acids. Each amino acid subunit is directed to its proper site in the protein chain by a sequence of code letters in the DNA molecule (or molecules) that each organism inherits from its ancestors. It is this DNA that is shaped by evolution. Organisms compete with each other for survival; occasional random changes in their information content, carried by DNA, are sometimes advantageous in this competition. In this way organisms slowly become enriched with instructions facilitating their survival.

EXPERIMENT BEGINS when cells of the colon bacillus are ground in a mortar with finely divided aluminum oxide. "Sap" released from ruptured cells still synthesizes protein.

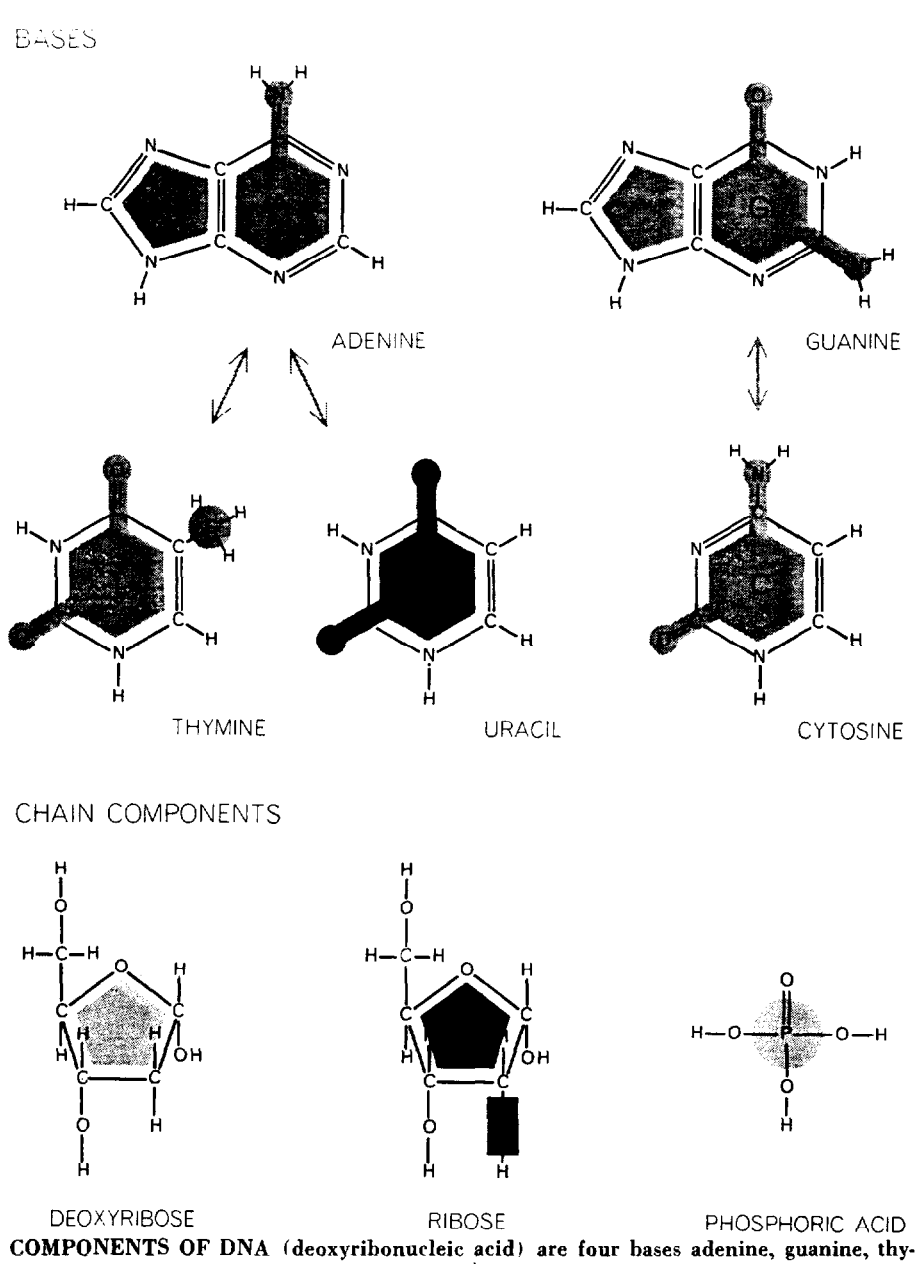
The exact number of proteins required for the functioning of a typical living cell is not known, but it runs to many hundreds. The great majority, if not all, of the proteins act as enzymes, or biological catalysts, which direct the hundreds of different chemical reactions that go on simultaneously within each cell. A typical protein is a molecular chain containing about 200 amino acid subunits linked together in a specific sequence. Each protein usually contains all or most of the 20 different kinds of amino acids. The code for each protein is carried by a single gene, which in turn is a particular region on the linear DNA molecule. To describe a protein containing 200 amino acid subunits a gene must contain at least 200 code words, represented by a sequence of perhaps 600 bases. No one yet knows the complete base sequence for a single gene. Viruses, the smallest structures containing the blueprints for their own replication, may contain from a few to several hundred genes. Bacteria may contain 1,000 genes; a human cell may contain a million. The human genes are not strung together in





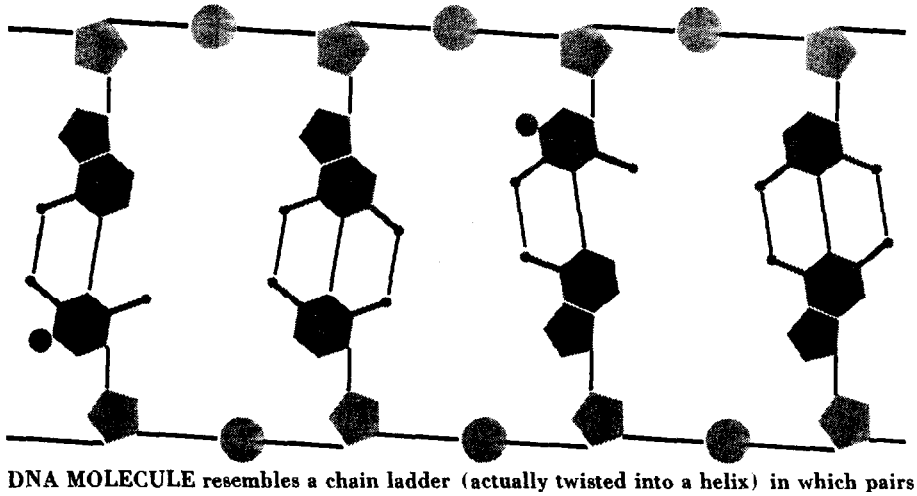


STEPS IN CODE BREAKING are shown in this sequence of photographs taken in the author's laboratory at the National Institutes of Health in Bethesda, Md. The open test tubes at upper left contain samples of the cell-free bacterial system capable of synthesizing protein when properly stimulated. The photograph shows stimulants being added. They include synthetic "messenger RNA" (ribonucleic acid) and amino acids, one of which is radioactive. The protein is produced when the samples are incubated 10 to 90 minutes. At upper right the protein is precipitated by the addition of trichloroacetic acid (TCA). At lower left the precipitate is transferred to filter-paper disks, which will be placed in carriers called planchettes. At lower right the planchettes are stacked in a radiation counting unit. Radiation measurement indicates how well a given sample of messenger RNA has directed amino acids into protein.



COMPONENTS OF DNA (deoxyribonucleic acid) are four bases adenine, guanine, thymine and cytosine (symbolized A, G, T, C), which act as code letters. Other components, deoxyribose and phosphoric acid, form chains to which bases attach (*see below*). In closely related RNA, uracil (U) replaces thymine and ribose replaces deoxyribose.

DNA STRUCTURE



DNA MOLECULE resembles a chain ladder (actually twisted into a helix) in which pairs of bases join two linear chains constructed from deoxyribose and phosphate subunits. The bases invariably pair so that A links to T and G to C. The genetic code is the sequence of bases as read down one side of the ladder. The deoxyribose-phosphate linkages in the two linear chains run in opposite directions. DNA molecules contain thousands of base pairs.

one long chain but must be divided among at least 46 DNA molecules. The minimum number is set by the number of human chromosomes (46), which collectively carry the hereditary material. In fact, each chromosome apparently carries not one or two but several copies of the same genetic message. If it were possible to assemble the DNA in a single human cell into one continuous thread, it would be about a yard long. This three-foot set of instructions for each individual is produced by the fusion of egg and sperm at conception and must be precisely replicated billions of times as the embryo develops.

The bottom illustration at left shows how the bases in DNA form the cross links connecting two helical strands composed of alternating units of deoxyribose (a simple sugar) and phosphate. The bases are attached to the sugar units and always occur in complementary pairs: A joined to T, and G joined to C. As a result one strand of the DNA molecule, with its associated bases, can serve as the template for creating a second strand that has a complementary set of bases. The faithful replication of genes during cell division evidently depends on such a copying mechanism.

The coding problem centers around the question: How can a four-letter alphabet (the bases A, G, C and T) specify a 20-word dictionary corresponding to the 20 amino acids? In 1954 the theoretical physicist George Gamow, now at the University of Colorado, pointed out that the code words in such a dictionary would have to contain at least three bases. It is obvious that only four code words can be formed if the words are only one letter in length. With two letters 4×4, or 16, code words can be formed. And with three letters $4 \times 4 \times 4$, or 64, code words become available—more than enough to handle the 20-word amino acid dictionary [*see top illustration on page 90*]. Subsequently many suggestions were made as to the nature of the genetic code, but extensive experimental knowledge of the code has been obtained only within the past 18 months.

The Genetic Messenger

It was recognized soon after the formulation of the Watson-Crick model of DNA that DNA itself might not be directly involved in the synthesis of protein, and that a template of RNA (ribonucleic acid) might be an intermediate in the process. Protein synthesis is conducted by cellular particles called ribosomes, which are about half protein and

half RNA (ribosomal RNA). Several years ago Jacques Monod and François Jacob of the Pasteur Institute in Paris coined the term "messenger RNA" to describe the template RNA that carried genetic messages from DNA to the ribosomes.

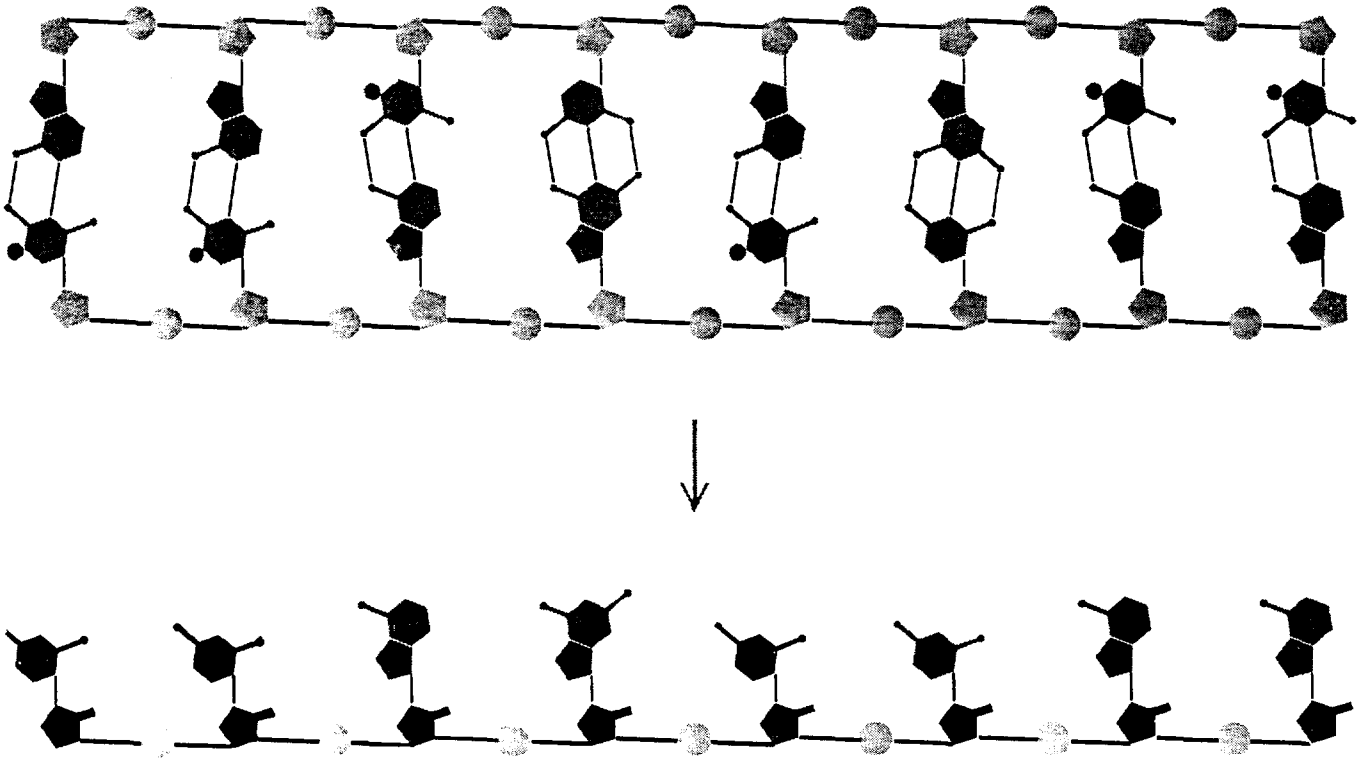
A few years ago evidence for the enzymatic synthesis of RNA complementary to DNA was found by Jerard Hurwitz of the New York University School of Medicine, by Samuel Weiss of the University of Chicago, by Audrey Stevens of St. Louis University and their respective collaborators [see "Messenger RNA," by Jerard Hurwitz and J. J. Furth; SCIENTIFIC AMERICAN, February, 1962]. These groups, and others, showed that an enzyme, RNA polymerase, catalyzes the synthesis of strands of RNA on the pattern of strands of DNA.

RNA is similar to DNA except that RNA contains the sugar ribose instead of deoxyribose and the base uracil instead of thymine. When RNA is being formed on a DNA template, uracil appears in the RNA chain wherever adenine appears at the complementary site on the DNA chain. One fraction of the RNA formed by this process is messenger RNA; it directs the synthesis of protein. Messenger RNA leaves the nucleus of the cell and attaches to the ribosomes. The sequence of bases in the messenger RNA specifies the amino acid sequence in the protein to be synthesized.

The amino acids are transported to the proper sites on the messenger RNA by still another form of RNA called transfer RNA. Each cell contains a specific activating enzyme that attaches a specific amino acid to its particular transfer RNA. Moreover, cells evidently contain more than one kind of transfer RNA capable of recognizing a given amino acid. The significance of this fact will become apparent later. Although direct recognition of messenger RNA code words by transfer RNA molecules has not been demonstrated, it is clear that these molecules perform at least part of the job of placing amino acids in the proper position in the protein chain. When the amino acids arrive at the proper site in the chain, they are linked to each other by enzymic processes that are only partly understood. The linking is accomplished by the formation of a peptide bond: a chemical bond created when a molecule of water is removed from two adjacent molecules of amino acid. The process requires a transfer enzyme, at least one other enzyme and a cofactor: guanosine triphosphate. It appears that amino acid subunits are bonded into the growing protein chain one at a time, starting at the end of the chain carrying an amino group (NH_2) and proceeding toward the end that terminates with a carboxyl group (COOH).

The process of protein synthesis can be studied conveniently in cell-free extracts of the colon bacillus (*Escherichia coli*). The bacteria grow rapidly in suitable nutrients and are harvested by sedimenting them out of suspension with a centrifuge. The cells are gently broken open by grinding them with finely powdered alumina [*see illustration on page 80*]; this releases the cell sap, containing DNA, messenger RNA, ribosomes, enzymes and other components. Such extracts are called cell-free systems, and when they are fortified with energy-rich substances (chiefly adenosine triphosphate), they readily incorporate amino acids into protein. The incorporation process can be followed by using amino acids containing carbon 14, a radioactive isotope of carbon.

Optimal conditions for protein synthesis in bacterial cell-free systems were determined by workers in many laboratories, notably Alfred Tissières of Harvard University, Marvin Lamborg and Paul C. Zamecnik of the Massachusetts General Hospital, G. David Novelli of the Oak Ridge National Laboratory and Sol Spiegelman of the University of Illinois. When we began our work at the National Institutes of Health, our

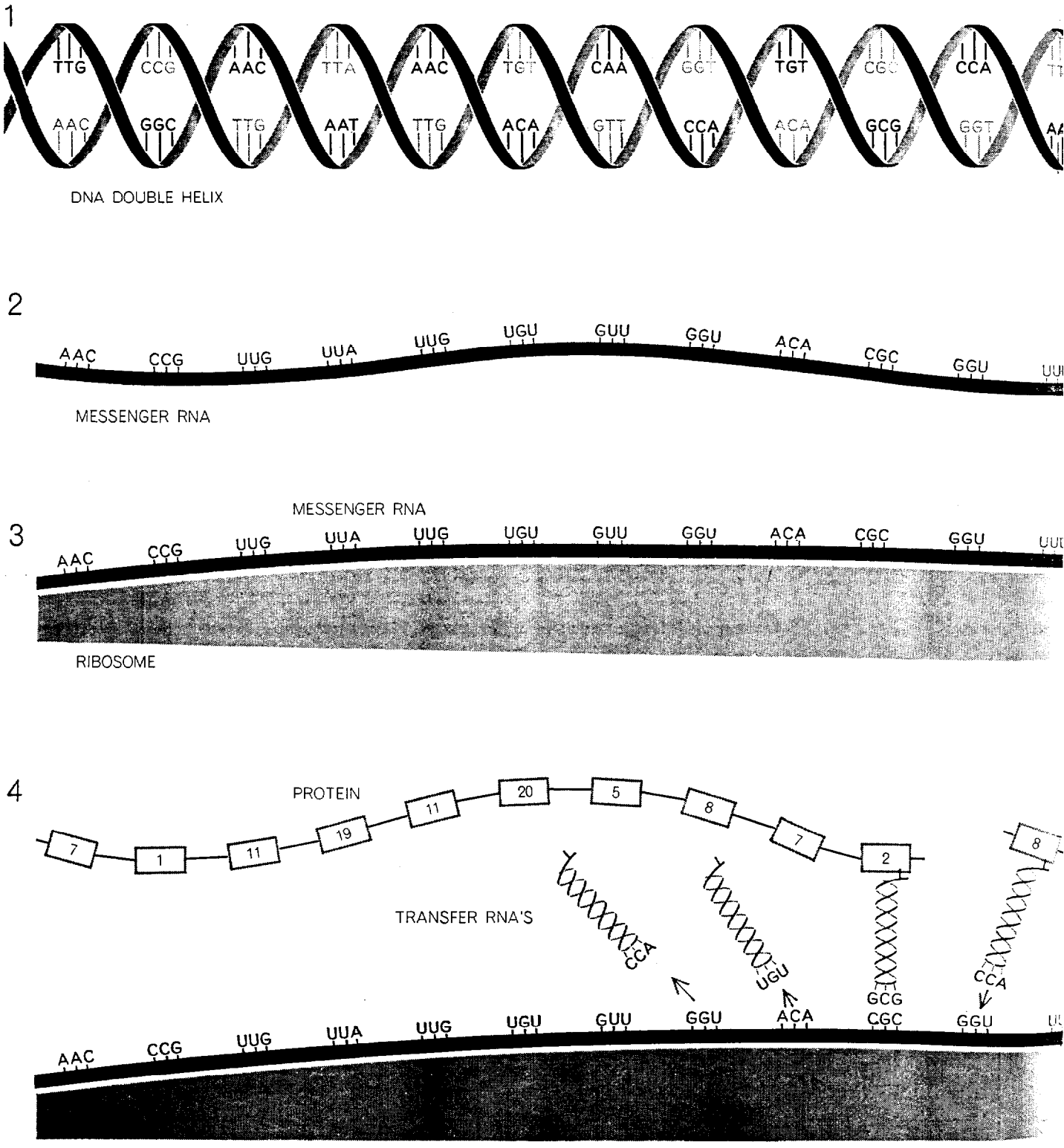


MESSENGER RNA is the molecular agent that transcribes the genetic code from DNA and carries it to the sites in the cell (the ribosomes) where protein synthesis takes place. The letters in messenger RNA are complementary to those in one strand of the DNA molecule. In this example UUAGUCAA is complementary to AATCAGTT. The exact mechanism of transcription is not known.

progress was slow because we had to prepare fresh enzyme extracts for each experiment. Later my colleague J. Heinrich Matthaei and I found a way to stabilize the extracts so that they could be stored for many weeks without appreciable loss of activity.

Normally the proteins produced in such extracts are those specified by the cell's own DNA. If one could establish the base sequence in one of the cell's genes—or part of a gene—and correlate it with the amino acid sequence in the protein coded by that gene, one would be able to translate the genetic code. Although the amino acid sequence is known for a number of proteins, no one has yet determined the base sequence of a gene, hence the correlation cannot be performed.

The study of cell-free protein syn-



SYNTHESIS OF PROTEIN begins with the genetic code embodied in DNA (*1*). The code is transcribed into messenger RNA (*2*). In the diagram it is assumed that the message has been derived from the DNA strand bearing dark letters. The messenger RNA finds its way to a ribosome (*3*), the site of protein synthesis. Amino acids, indicated by numbered rectangles, are carried to proper sites on the messenger RNA by molecules of transfer RNA (*see illustration on opposite page*). Bases are actually equidistant, not

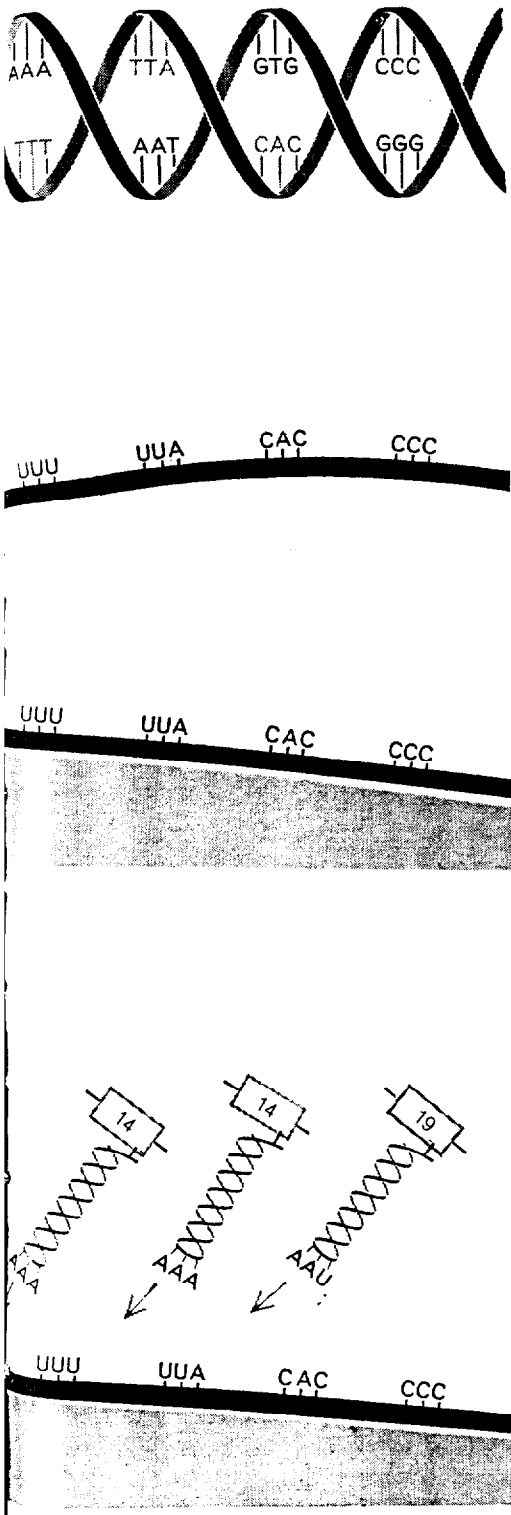
thesis provided an indirect approach to the coding problem. Tissières, Novelli and Bernard Nisman, then at the Pasteur Institute, had reported that protein synthesis could be halted in cell-free extracts by adding deoxyribonuclease, or DNAase, an enzyme that specifically destroys DNA. Matthaei and I also observed this effect and studied its characteristics. It seemed probable that protein synthesis stopped after the messenger RNA had been depleted. When we added crude fractions of messenger RNA to such extracts, we found that they stimulated protein synthesis. The development of this cell-free assay for messenger RNA provided the rationale for all our subsequent work.

We obtained RNA fractions from various natural sources, including viruses, and found that many of them were highly active in directing protein synthesis in the cell-free system of the colon bacillus. The ribosomes of the colon bacillus were found to accept RNA "blueprints" obtained from foreign organisms, including viruses. It should be emphasized that only minute amounts of protein were synthesized in these experiments.

It occurred to us that synthetic RNA containing only one or two bases might direct the synthesis of simple proteins containing only a few amino acids. Synthetic RNA molecules can be prepared with the aid of an enzyme, polynucleotide phosphorylase, found in 1955 by Marianne Grunberg-Manago and Severo Ochoa of the New York University School of Medicine. Unlike RNA polymerase, this enzyme does not follow the pattern of DNA. Instead it forms RNA polymers by linking bases together in random order.

A synthetic RNA polymer containing only uracil (called polyuridylic acid, or poly-U) was prepared and added to the active cell-free system together with mixtures of the 20 amino acids. In each mixture one of the amino acids contained radioactive carbon 14; the other 19 amino acids were nonradioactive. In this way one could determine the particular amino acid directed into protein by poly-U.

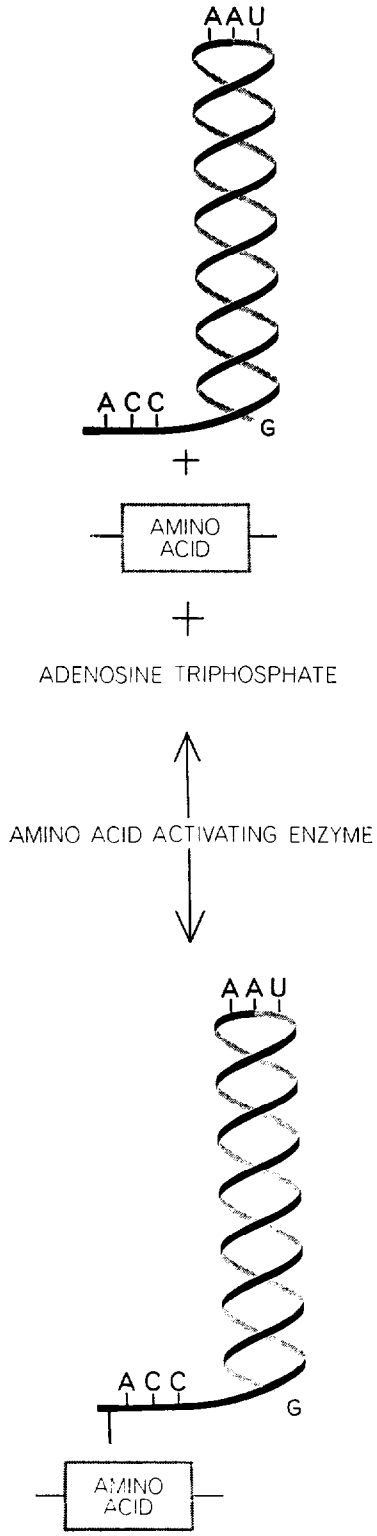
It proved to be the amino acid phenylalanine. This provided evidence that the RNA code word for phenylalanine was a sequence of U's contained in poly-U. The code word for another amino acid, proline, was found to be a sequence of C's in polycytidylic acid, or poly-C. Thus a cell-free system capable of synthesizing protein under the direction of chemically defined preparations of RNA provided a simple means for translating the genetic code.



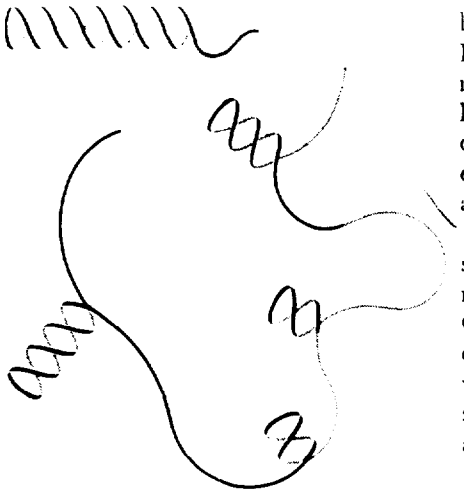
grouped in triplets, and mechanism of recognition between transfer RNA and messenger RNA is hypothetical. Linkage of amino acid subunits creates a protein molecule.

The Code-Word Dictionary

Ochoa and his collaborators and our group at the National Institutes of



TRANSFER RNA is a special helical form of RNA that transports amino acids to their proper site in the protein chain. There is at least one transfer RNA for each of the 20 common amino acids. All, however, seem to carry the bases ACC where the amino acids attach and G at the opposite end. The attachment requires a specific enzyme and energy supplied by adenosine triphosphate. Unpaired bases in transfer RNA (AAU in the example) may provide the means by which the transfer RNA "recognizes" the place to deposit its amino acid package.



RNA STRUCTURE can take various forms. Transfer RNA (*top*) seems to be a fairly short double helix (probably less perfect than shown) that is closed at one end. Some RNA molecules contain a mixture of coiled and uncoiled regions (*bottom*).

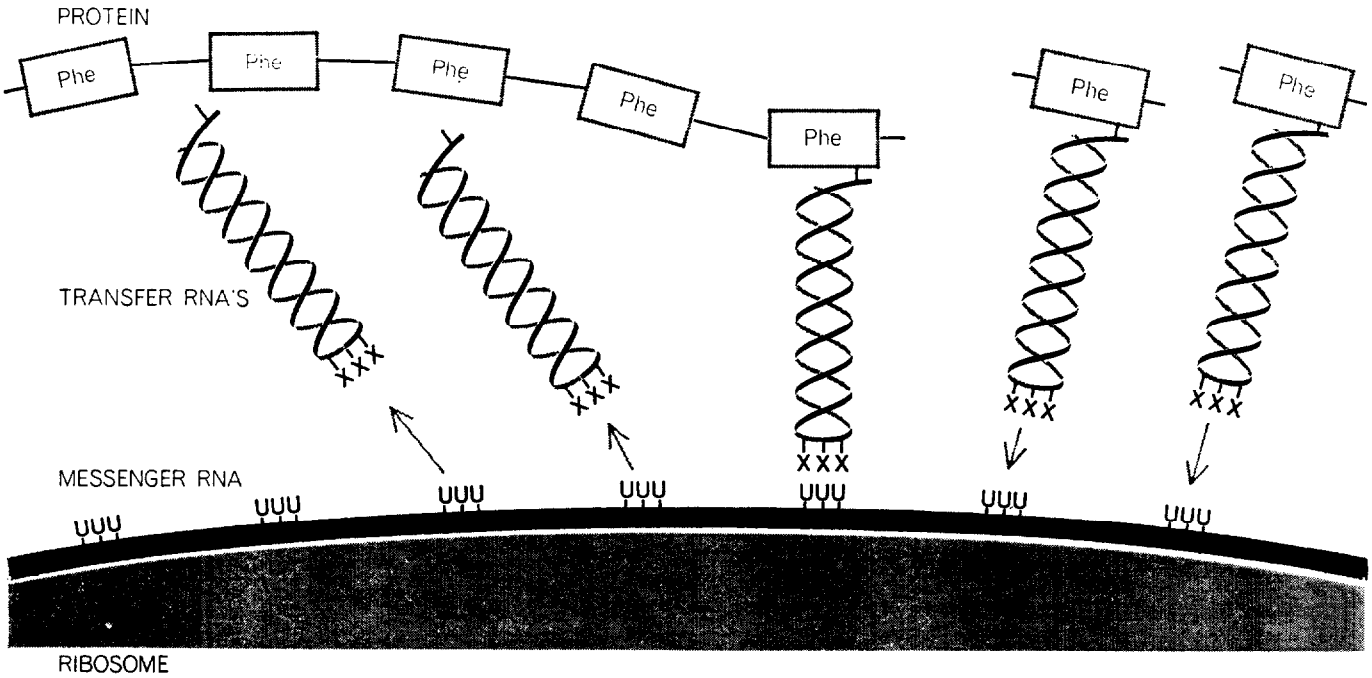
Health, working independently, have now synthesized and tested polymers containing all possible combinations of the four RNA bases A, G, C and U. In the initial experiments only RNA polymers containing U were assayed, but recently many non-U polymers with high template activity have been found by M. Bretscher and Grunberg-Manago of the University of Cambridge, and also by Oliver W. Jones and me. All the results so far are summarized in the table at the bottom of pages 90 and 91. It lists the RNA polymers containing the minimum number of bases capable of stimulating protein formation. The inclusion of another base in a polymer usually enables it to code for additional amino acids.

With only two kinds of base it is possible to make six varieties of RNA polymer: poly-AC, poly-AG, poly-AU, poly-CG, poly-CU and poly-CU. If the ratio of the bases is adjusted with care, each variety can be shown to code with great specificity for different sets of amino acids. The relative amount of one amino acid directed into protein compared with another depends on the ratio of bases in the RNA. Assuming a random sequence of bases in the RNA, the theoretical probabilities of finding particular sequences of two, three or more bases can be calculated easily if the base ratio is known. For example, if poly-UC contains 70 per cent U and 30 per cent C, the probability of the occurrence of the triplet sequence UUU is $.7 \times .7 \times .7$, or .34. That is, 34 per cent of the triplets in the polymer are expected to be UUU. The probability of obtaining the sequence UUC is $.7 \times .7 \times .3$, or .147. Thus 14.7 per cent of the triplets in such a polymer are probably UUC. This type of calculation, however, assumes randomness, and it is not certain that all the actual polymers are truly random.

It had been predicted by Gamow, Crick and others that for each amino acid there might be more than one code word, since there are 64 possible triplets and only 20 amino acids. A code with multiple words for each object coded is termed degenerate. Our experiments show that the genetic code is indeed degenerate. Leucine, for example, is coded by RNA polymers containing U alone, or U and A, or U and C, or U and G.

It must be emphasized that degeneracy of this sort does not imply lack of specificity in the construction of proteins. It means, rather, that a specific amino acid can be directed to the proper site in a protein chain by more than one code word. Presumably this flexibility of coding is advantageous to the cell in ways not yet fully understood.

A molecular explanation of degeneracy has been provided recently in a striking manner. It has been known that some organisms contain more than one species of transfer RNA capable of recognizing a given amino acid. The colon bacillus, for example, contains two readily distinguishable species that transfer leucine. Bernard Weisblum and Seymour Benzer of Purdue University and Robert W. Holley of Cornell University separated the two leucine-transfer species and tested them in cell-free systems. They found that one of the species recognizes poly-UC but not poly-UG. The other species recognizes poly-UG but not poly-UC [*see top illustration on page 89*]. Although the number of transfer RNA species per cell is unknown, it is possible



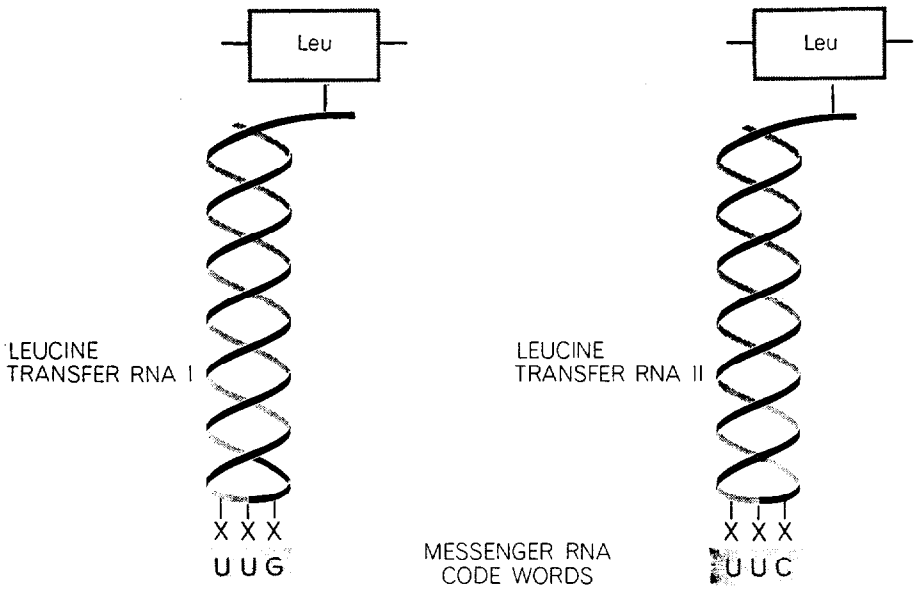
FIRST BREAK IN GENETIC CODE was the discovery that a synthetic messenger RNA containing only uracil (poly-U) directed the manufacture of a synthetic protein containing only one amino acid, phenylalanine (*Phe*). The finding was made by the author and J. Heinrich Matthaei. The X's in transfer RNA signify that the bases that respond to code words in messenger RNA are not known.

that each species corresponds to a different code word.

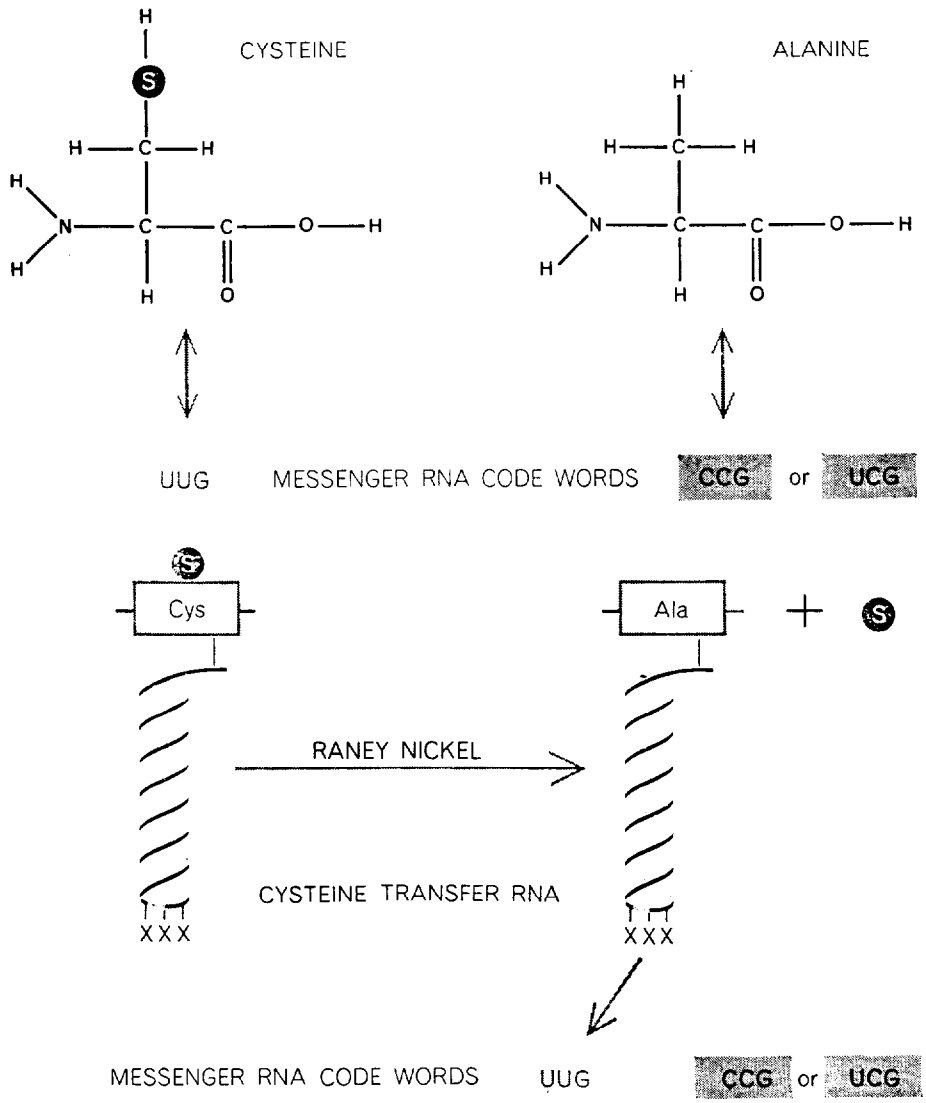
There is, however, the possibility of real ambiguity in protein synthesis. This would occur if one code word were to direct two or more kinds of amino acid into protein. So far only one such ambiguity has been found. Poly-U directs small amounts of leucine as well as phenylalanine into protein. The ratio of the two amino acids incorporated is about 20 or 30 molecules of phenylalanine to one of leucine. In the absence of phenylalanine, poly-U codes for leucine about half as well as it does for phenylalanine. The molecular basis of this ambiguity is not known. Nor is it known if the dual coding occurs in living systems as well as in cell-free systems.

Base sequences that do not encode for any amino acid are termed "nonsense words." This term may be misleading, for such sequences, if they exist, might have meaning to the cell. For example, they might indicate the beginning or end of a portion of the genetic message. An indirect estimate of the frequency of nonsense words can be obtained by comparing the efficiency of random RNA preparations with that of natural messenger RNA. We have found that many of the synthetic polymers containing four, three or two kinds of base are as efficient in stimulating protein synthesis as natural polymers are. This high efficiency, together with high coding specificity, suggests that relatively few base sequences are nonsense words.

In his recent article in *Scientific American* Crick presented arguments for believing that the coding ratio is either three or a multiple of three. Recently we have determined the relative amounts of different amino acids directed into protein by synthetic RNA preparations of known base ratios, and the evidence suggests that some code words almost surely contain three bases. Yet, as the table at the bottom of the next two pages shows, 18 of the 20 amino acids can be coded by words containing only two different bases. The exceptions are aspartic acid and methionine, which seem to require some combination of U, G and A. (Some uncertainty still exists about the code words for these amino acids, because even poly-UGA directs very little aspartic acid or methionine into protein.) If the entire code indeed consists of triplets, it is possible that correct coding is achieved, in some instances, when only two out of the three bases read are recognized. Such imperfect recognition might occur more often with synthetic RNA polymers containing



TWO KINDS OF TRANSFER RNA have been found, each capable of transporting leucine (*Leu*). One kind (*left*) recognizes the code word UUG; the other (*right*) recognizes UUC.



INGENIOUS EXPERIMENT showed that code-word recognition depends on the specificity of transfer RNA, not on the structure of the amino acid being transported. Cysteine is coded by UUG, alanine by CCG or UCG. Cysteine was hooked to its specific transfer RNA and sulfur was removed by a catalyst (Raney nickel). With sulfur removed from the molecule, cysteine became alanine. It was still directed into protein, however, as if it were cysteine.

SINGLET CODE (4 WORDS)	DOUBLET CODE (16 WORDS)				TRIPLET CODE (64 WORDS)			
A	AA	AG	AC	AU	AAA	AAG	AAC	AAU
G	GA	GG	GC	GU	AGA	AGG	AGC	AGU
C	CA	CG	CC	CU	ACA	ACG	ACC	ACU
U	UA	UG	UC	UU	AUA	AUG	AUC	AUU
					GAA	GAG	GAC	GAU
					GGA	GGG	GGC	GGU
					GCA	GCG	GCC	GCU
					GUA	GUG	GUC	GUU
					CAA	CAG	CAC	CAU
					CGA	CGG	CGC	CGU
					CCA	CCG	CCC	CCU
					CUA	CUG	CUC	CUU
					UAA	UAG	UAC	UAU
					UGA	UGG	UGC	UGU
					UCA	UCG	UCC	UCU
					UUA	UUG	UUC	UUU

CODE-LETTER COMBINATIONS increase sharply with the length of the code word. Since at least 20 code words are needed to identify the 20 common amino acids, the minimum code length is a sequence of three letters, assuming that all words are the same length.

only one or two bases than it does with natural messenger RNA, which always contains a mixture of all four. The results obtained with synthetic RNA may demonstrate the coding potential of the cell; that is, it may reveal code words that function routinely in the living cell and potential words that would be recognized if appropriate mutations were to occur in the cellular DNA. The table on page 93 summarizes the code-word dictionary on the assumption that all code words are triplets.

The Universality of the Code

Does each plant or animal species have its own genetic code, or is the same genetic language used by all species on this planet? Preliminary evidence suggests that the code is essentially universal and that even species at opposite ends of the evolutionary scale use much the same code. For instance, a number of laboratories in the U.S. and England have recently reported that synthetic RNA polymers code the same way in mammalian cell-free systems as they do in the bacterial system. The base compositions of mammalian code words corresponding to about six amino acids have been determined so far. It nevertheless seems probable that some differences may be found in the future. Since certain amino acids are coded by multiple words, it is not unlikely that one species may use one word and another species a different one.

An indirect check on the validity of code words obtained in cell-free systems can be made by studying natural proteins that differ in amino acid composition at only one point in the protein chain. For example, the hemoglobin of an individual suffering from "sickle cell" anemia differs from normal hemoglobin in that it has valine at one point in the chain instead of glutamic acid. Another abnormal hemoglobin has lysine at the same point. One might be able to show, by examining the code-word dictionary, that these three amino acids—glutamic acid, valine and lysine—have similar code words. One could then infer that the two abnormal hemoglobins came into being as a result of a mutation that substituted a single base for another in the gene that controls the production of hemoglobin. As a matter of fact, the code-word dictionary shows that the code words are similar enough for this to have happened. One of the code groups for glutamic acid is AGU. Substitution of a U for A produces UGU, the code group for valine. Substitution of an A for a U yields AGA, one of the code groups for lysine. Similar analyses have been made for other proteins in which amino acid substitutions are known, and in most cases the substitutions can be explained by alteration of a single base in code-word triplets. Presumably more code words will be found in the future and the correlation between genetic base sequences and amino acid sequences can be made with greater assurance.

The Nature of Messenger RNA

Does each molecule of messenger RNA function only once or many times in directing the synthesis of protein? The question has proved difficult because most of the poly-U in the experimental system is degraded before it is able to function as a messenger. We have found, nevertheless, that only about 1.5 U's in poly-U are required to direct the incorporation of one molecule of phenylalanine into protein. And George Spyrides and Fritz A. Lipmann of the Rockefeller Institute have reported that only about .75 U's are required per molecule of amino acid in their studies. If the coding is done by triplets, three U's would be

AMINO ACIDS CODED	U	A	C	G
	PHENYLALANINE	LYSINE	PROLINE●	
	LEUCINE■			

■ POLY U CODES PREFERENTIALLY FOR PHENYLALANINE
● REPORTED BY ONLY ONE LABORATORY; STILL TO BE CONFIRMED
▲ REQUIRES ONLY FIRST OF TWO BASES LISTED
△ REQUIRES ONLY SECOND OF TWO BASES LISTED

SPECIFICITY OF CODING is shown in this table, which lists 18 amino acids that can be coded by synthetic RNA polymers containing no more than one or two kinds of base. The only amino acids that seem to require more than two bases for coding are aspartic acid and methionine, which need U, A and G. The relative amounts of amino acids directed into pro-

required if the messenger functioned only once. Evidently each poly-U molecule directs the synthesis of more than one long-chain molecule of polyphenylalanine. Similar results have been obtained in intact cells. Cyrus Levinthal and his associates at the Massachusetts Institute of Technology inhibited messenger RNA synthesis in living bacteria with the antibiotic actinomycin and found that each messenger RNA molecule present at the time messenger synthesis was turned off directed the synthesis of 10 to 20 molecules of protein.

We have observed that two factors in addition to base sequence have a profound effect on the activity of messenger RNA: the length of the RNA chain and its over-all structure. Poly-U molecules that contain more than 100 U's are much more active than molecules with fewer than 50. Robert G. Martin and Bruce Ames of the National Institutes of Health have found that chains of poly-U containing 450 to 700 U's are optimal for directing protein synthesis.

There is still much to be learned about the effect of structure on RNA function. Unlike DNA, RNA molecules are usually single-stranded. Frequently, however, one part of the RNA molecule loops back and forms hydrogen bonds with another portion of the same molecule. The extent of such internal pairing is influenced by the base sequence in the molecule. When poly-U is in solution, it usually has little secondary structure; that is, it consists of a simple chain with few, if any, loops or knots. Other types of RNA molecules display a considerable amount of secondary structure [*see top illustration on page 86*].

We have found that such a secondary structure interferes with the activity of messenger RNA. When solutions of poly-U and poly-A are mixed, they form double-strand (U-A) and triple-strand (U-A-U) helices, which are completely inactive in directing the synthesis of polyphenylalanine. In collaboration with Maxine F. Singer of the National Institutes of Health we have shown that poly-UG containing a high degree of ordered secondary structure (possibly due to G-G hydrogen-bonding) is unable to code for amino acids.

It is conceivable that natural messenger RNA contains at intervals short regions of secondary structure resembling knots in a rope. These regions might signify the beginning or the end of a protein. Alternative hypotheses suggest that the beginning and end are indicated by particular base sequences in the genetic message. In any case it seems probable that the secondary structure assumed by different types of RNA will be found to have great influence on their biological function.

The Reading Mechanism

Still not completely understood is the manner in which a given amino acid finds its way to the proper site in a protein chain. Although transfer RNA was found to be required for the synthesis of polyphenylalanine, the possibility remained that the amino acid rather than the transfer RNA recognized the code word embodied in the poly-U messenger RNA.

To distinguish between these alternative possibilities, a brilliant experiment was performed jointly by François Chapeville and Lipmann of the Rockefeller Institute, Günter von Ehrenstein of Johns Hopkins University and three Purdue workers: Benzer, Weisblum and William J. Ray, Jr. One amino acid, cysteine, is directed into protein by poly-UG. Alanine, which is identical with cysteine except that it lacks a sulfur atom, is directed into protein by poly-CG or poly-UCG. Cysteine is transported by one species of transfer RNA and alanine by another. Chapeville and his associates enzymatically attached cysteine, labeled with carbon 14, to its particular type of transfer RNA. They then exposed the molecular complex to a nickel catalyst, called Raney nickel, that removed the sulfur from cysteine and converted it to alanine—without detaching it from cysteine-transfer RNA. Now they could ask: Will the labeled alanine be coded as if it were alanine or cysteine? They found it was coded by poly-UG, just as if it were cysteine [*see bottom illustration on page 89*]. This experiment shows that an amino acid loses its identity after combining with transfer RNA and is carried willy-nilly to the code word recognized by the transfer RNA.

The secondary structure of transfer RNA itself has been clarified further this past year by workers at King's College of the University of London. From X-ray evidence they have deduced that transfer RNA consists of a double helix very much like the secondary structure found in DNA. One difference is that the transfer RNA molecule is folded back on itself, like a hairpin that has been twisted around its long axis. The molecule seems to contain a number of unpaired bases; it is possible that these provide the means for recognizing specific code words in messenger RNA [*see illustration at right on page 85*].

There is still considerable mystery about the way messenger RNA attaches to ribosomes and the part that ribosomes play in protein synthesis. It has been known for some time that colon bacillus ribosomes are composed of at least two types of subunit and that under certain conditions they form aggregates consisting of two subunits (dimers) and four subunits (tetramers). In collaboration with Samuel Barondes, we found

BASES PRESENT IN SYNTHETIC RNA

UA	UC	UG	AC	AG	CG
PHENYLALANINE ▲	PHENYLALANINE ▲	PHENYLALANINE ▲	LYSINE ▲	LYSINE ▲	PROLINE ▲
LYSINE ▲	PROLINE ▲	LEUCINE	PROLINE ▲	GLUTAMIC ACID	ARGININE ●
TYROSINE	LEUCINE	VALINE	HISTIDINE	ARGININE ●	ALANINE ●
LEUCINE	SERINE	CYSTEINE	ASPARAGINE	GLUTAMINE ●	
ISOLEUCINE		TRYPTOPHAN	GLUTAMINE	GLYCINE ●	
ASPARAGINE ●		GLYCINE	THREONINE		

tein by RNA polymers containing two bases depend on the base ratios. When the polymers contain a third and fourth base, additional kinds of amino acids are incorporated into protein. Thus the activity of poly-UCG (an RNA polymer containing U, C and G) resembles that of poly-UC plus poly-UG. Poly-G has not been found to code for any amino acid. Future work will undoubtedly yield data that will necessitate revisions in this table. An RNA-code-word dictionary derived from the table appears on page 93.

that the addition of poly-U to reaction mixtures initiated further ribosome aggregation. In early experiments only tetramers or still larger aggregates supported the synthesis of polyphenylalanine. Spyrides and Lipmann have shown that poly-U makes only certain "active" ribosomes aggregate and that the remaining monomers and dimers do not support polyphenylalanine synthesis.

A possibly related phenomenon has been observed in living cells by Alexander Rich and his associates at the Massachusetts Institute of Technology. They find that in reticulocytes obtained from rabbit blood, protein synthesis seems to be carried out predominantly by aggregates of five ribosomes, which may be held together by a single thread of messenger RNA. They have named the aggregate a polysome.

Many compelling problems still lie ahead. One is to establish the actual sequence of bases in code words. At present the code resembles an anagram. We know the letters but not the order of most words.

Another intriguing question is whether in living cells the double strand of DNA serves as a template for the production of a single strand of messenger RNA, or whether each strand of DNA serves as a template for the production of two different, complementary strands of RNA. If the latter occurs—and available evidence suggests that it does—the function of each strand must be elucidated.

Ultimately one hopes that cell-free

AMINO ACID	RNA CODE WORDS			
ALANINE	CCG	UCG■		
ARGININE	CGC	AGA	UCG■	
ASPARAGINE	ACA	AUA		
ASPARTIC ACID	GUA			
CYSTEINE	UUG▲			
GLUTAMIC ACID	GAA	AGU■		
GLUTAMINE	ACA	AGA	AGU■	
GLYCINE	UGG	AGG		
HISTIDINE	ACC			
ISOLEUCINE	UAU	UAA		
LEUCINE	UUG	UUC	UUA	UUU□
LYSINE	AAA	AAG●	AAU●	
METHIONINE	UGA■			
PHENYLALANINE	UUU			
PROLINE	CCC	CCU▲	CCA▲	CCG▲
SERINE	UCU	UCC	UCG	
THREONINE	CAC	CAA		
TRYPTOPHAN	GGU			
TYROSINE	AUU			
VALINE	UGU			

▲ UNCERTAIN WHETHER CODE IS UUG OR GGU

■ NEED FOR U UNCERTAIN

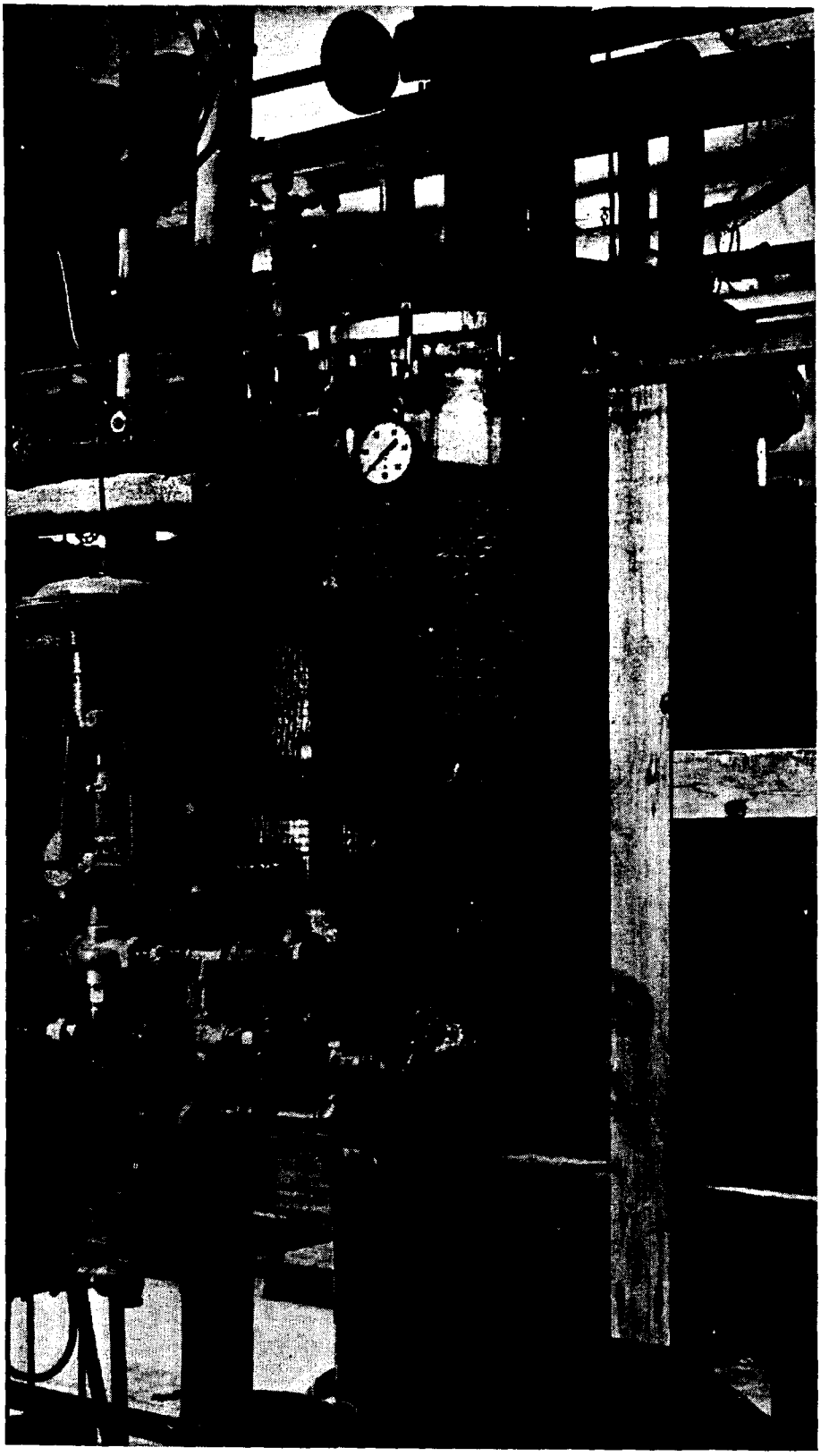
□ CODES PREFERENTIALLY FOR PHENYLALANINE

● NEED FOR G AND U UNCERTAIN

▲ NEED FOR U A G UNCERTAIN

GENETIC-CODE DICTIONARY lists the code words that correspond to each of the 20 common amino acids, assuming that all the words are triplets. The sequences of the letters in the code words have not been established, hence the order shown is arbitrary. Although half of the amino acids have more than one code word, it is believed that each triplet codes uniquely for a particular amino acid. Thus various combinations of AAC presumably code for asparagine, glutamine and threonine. Only one exception has been found to this presumed rule. The triplet UUU codes for phenylalanine and, less effectively, for leucine.

systems will shed light on genetic control mechanisms. Such mechanisms, still undiscovered, permit the selective retrieval of genetic information. Two cells may contain identical sets of genes, but certain genes may be turned on in one cell and off in another in highly specific fashion. With cell-free systems the powerful tools of enzymology can be brought to bear on these and other problems, with the promise that the molecular understanding of genetics will continue to advance rapidly in the near future.



PRODUCTION OF BACTERIA is carried out on a large scale at the National Institutes of Health. The vessel holds colon bacilli that the author and others need for experiments.